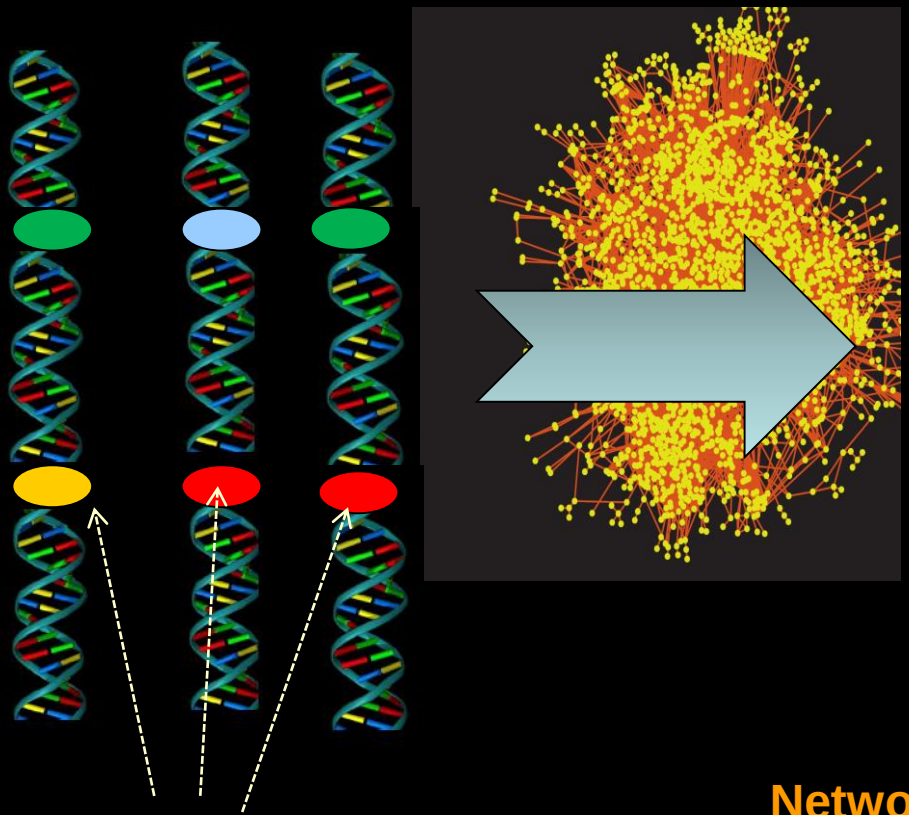
Network Biology Approach to Complex Diseases

Teresa Przytycka

NIH / NLM / NCBI

Recap: Genotype – Phenotype relation

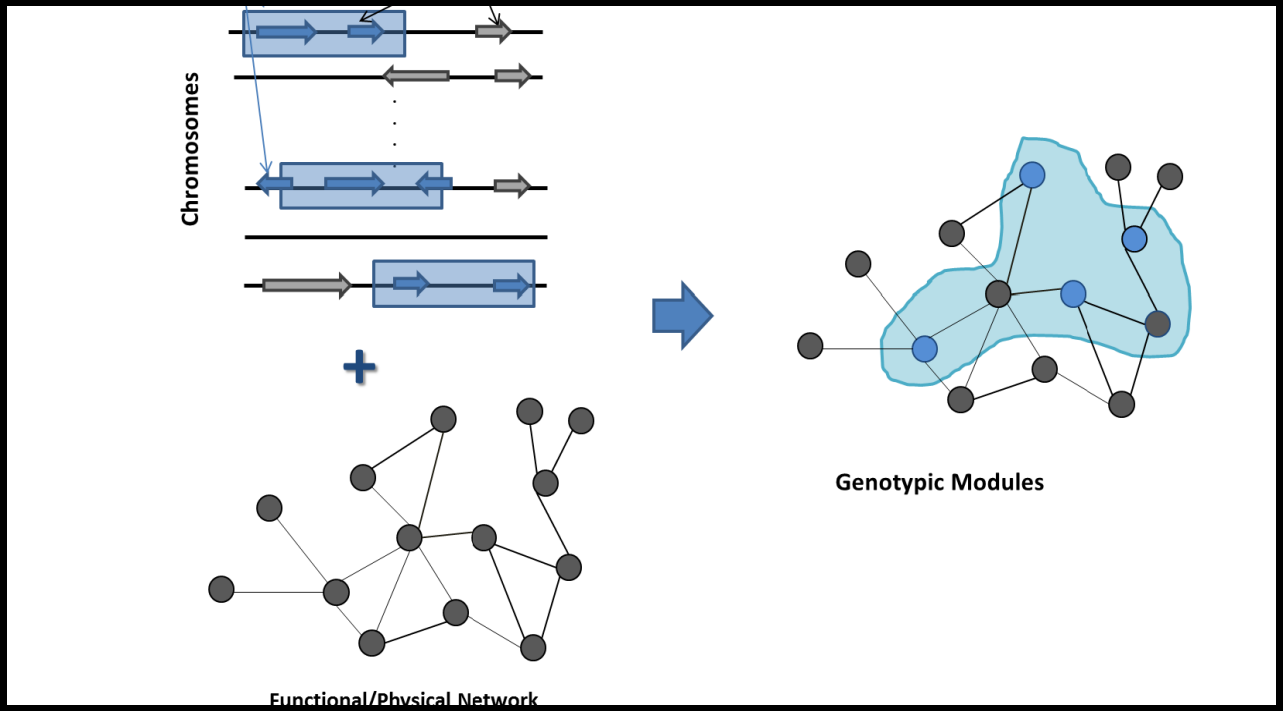
Individuals (genotype)



Individuals (phenotype)

Network based approaches – bringing knowledge of relation between molecules gained from high throughput experiments

Recap: Genotypic modules



Searching for genotypic modules:

- identification of genes/genomic regions that are frequently altered in a disease of interest
- mapping the genes residing in the altered regions to a network
- modules or subnetworks enriched with the altered genes are identified

Individual approaches differ in the way this last step is preformed

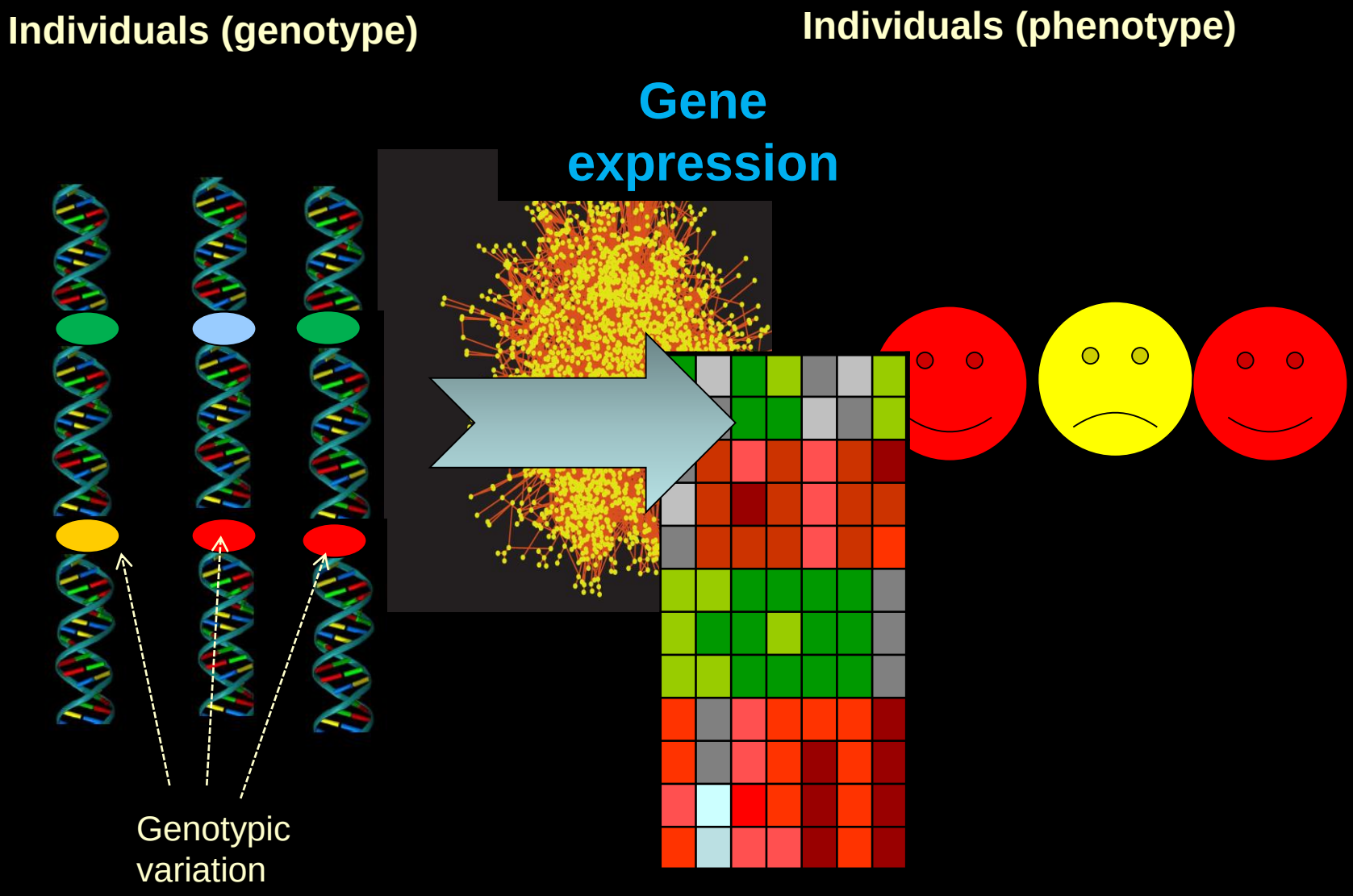
Network Biology Approach to Complex Diseases

LECTURE 2
Phenotypic / expression based dys-regulated modules
Combining expression and genetic data

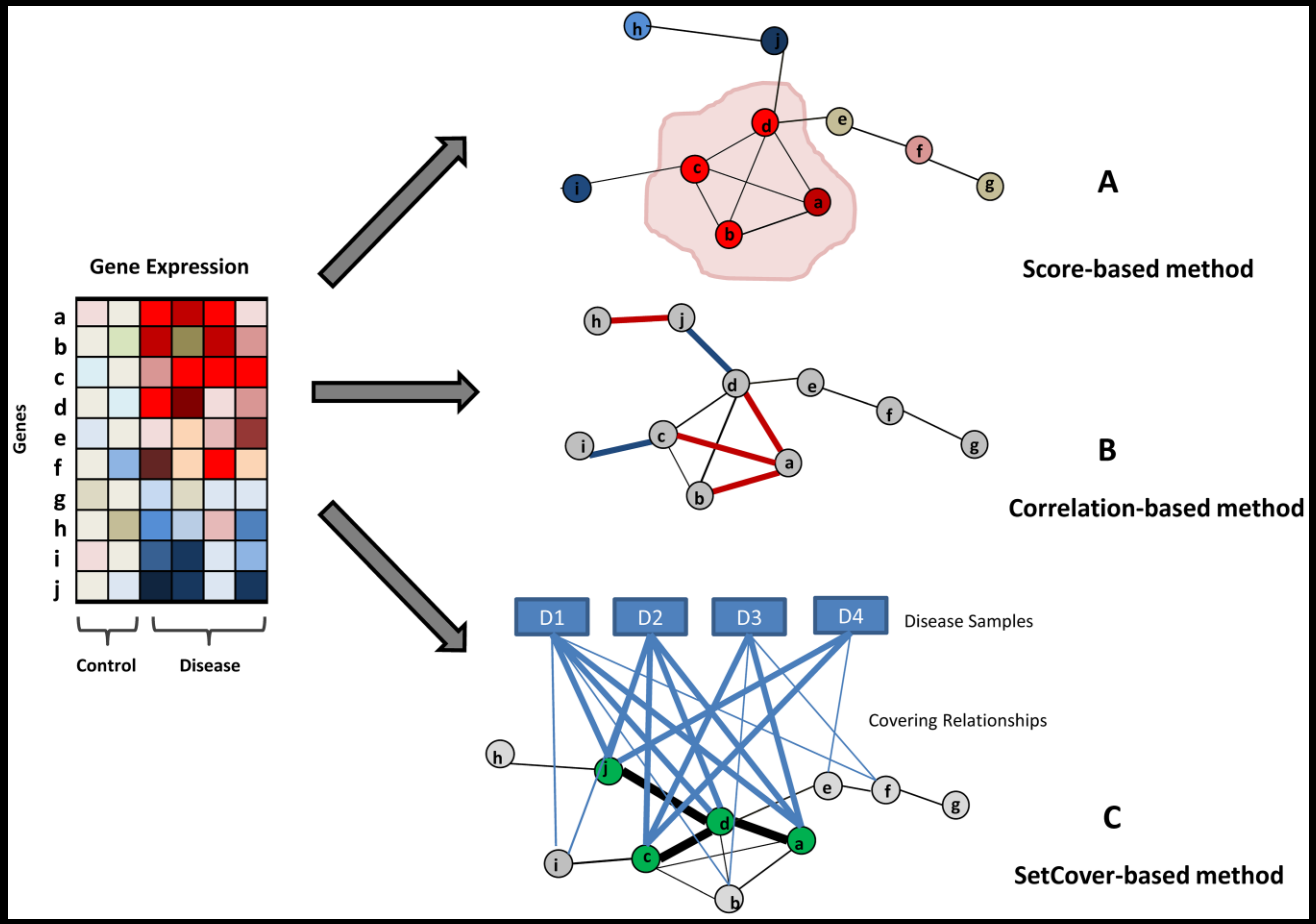
Teresa Przytycka

NIH / NLM / NCBI

Genotype – Phenotype relation



Main classes of methods to find expression dysregulated modules

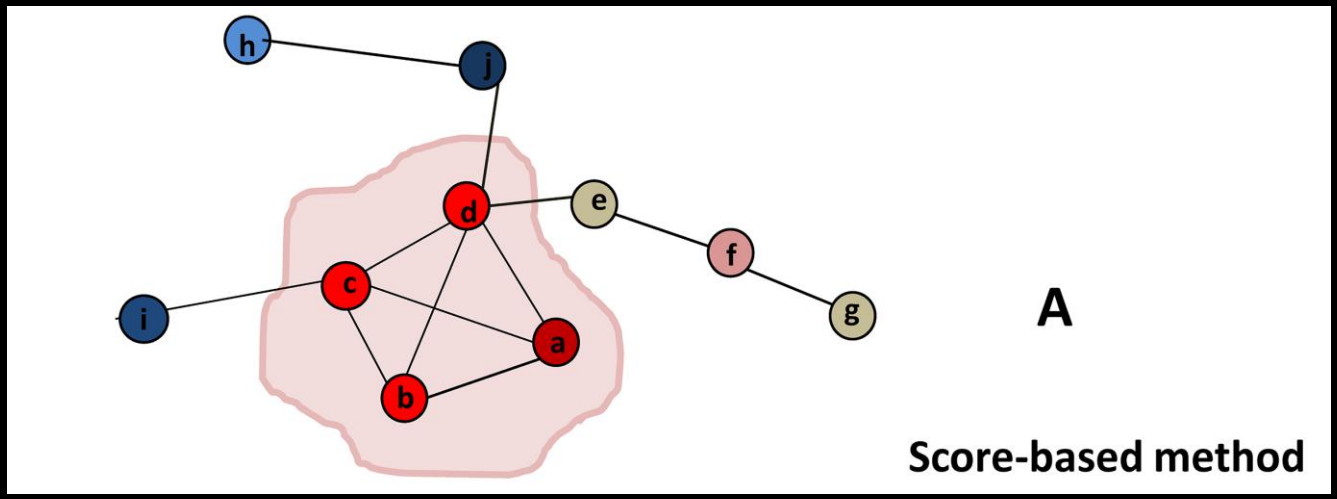


Why this is any different than genotypic modules ?

- Unlike genetic perturbation, expression changes are expected to affect direct neighbors thus you might be expectation a more continuous propagation of perturbation

- For the genetic perturbations we mapped the perturbed genes on the network but we had to fill the rest of the module based on our model – shortest path/Steiner tree/ heat diffusion

GROUP 1: Scoring based methods



Idea

- Score genes for differential expression
  - Given population of disease and normal used a statistical test for difference
  - Alternatively – use fold expression difference if not enoug samples
- Identify subnetworks with maximum total score

Example: Chuang et al. Network based classification of breast cancer metastasis

MosSysBio 2007

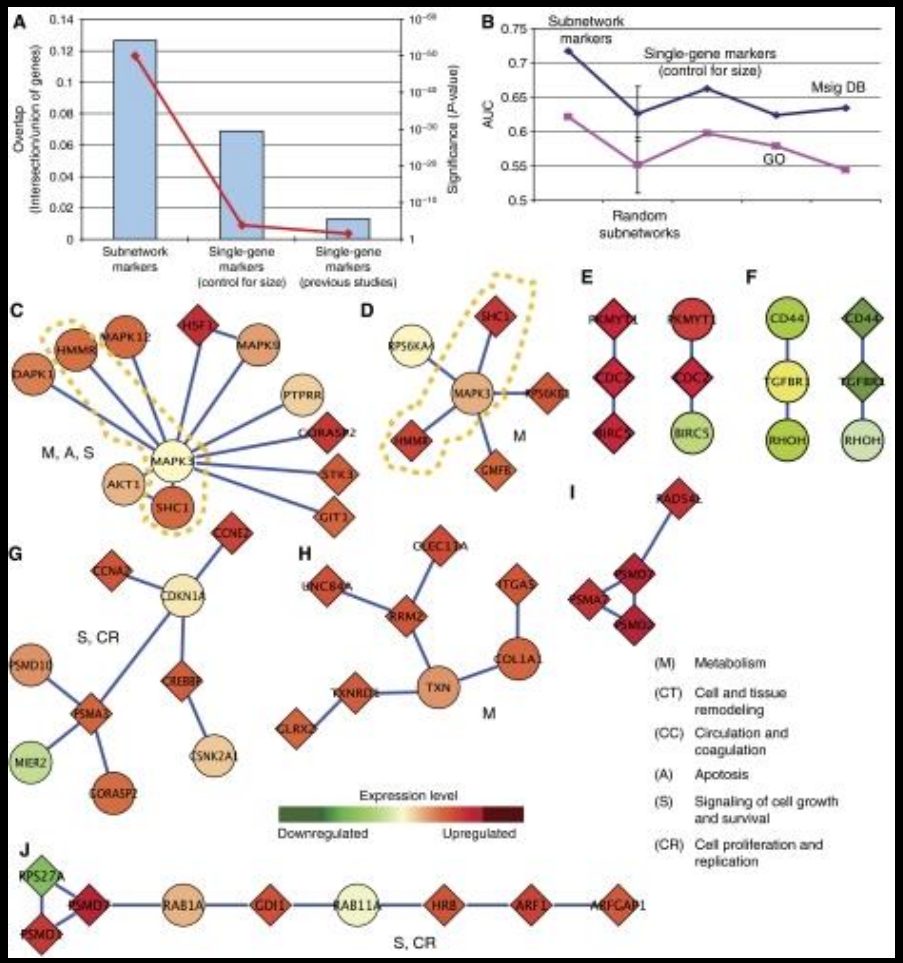
- Score how well each gene discriminates between case and control (in their case metastasis and non-metastasis)
- Score candidate subnetworks based on aggregate discriminative score
- Search for most discriminative subnetworks (greedy search)

Example: Chuang et al. Network based classification of breast cancer metastasis

MosSysBio 2007

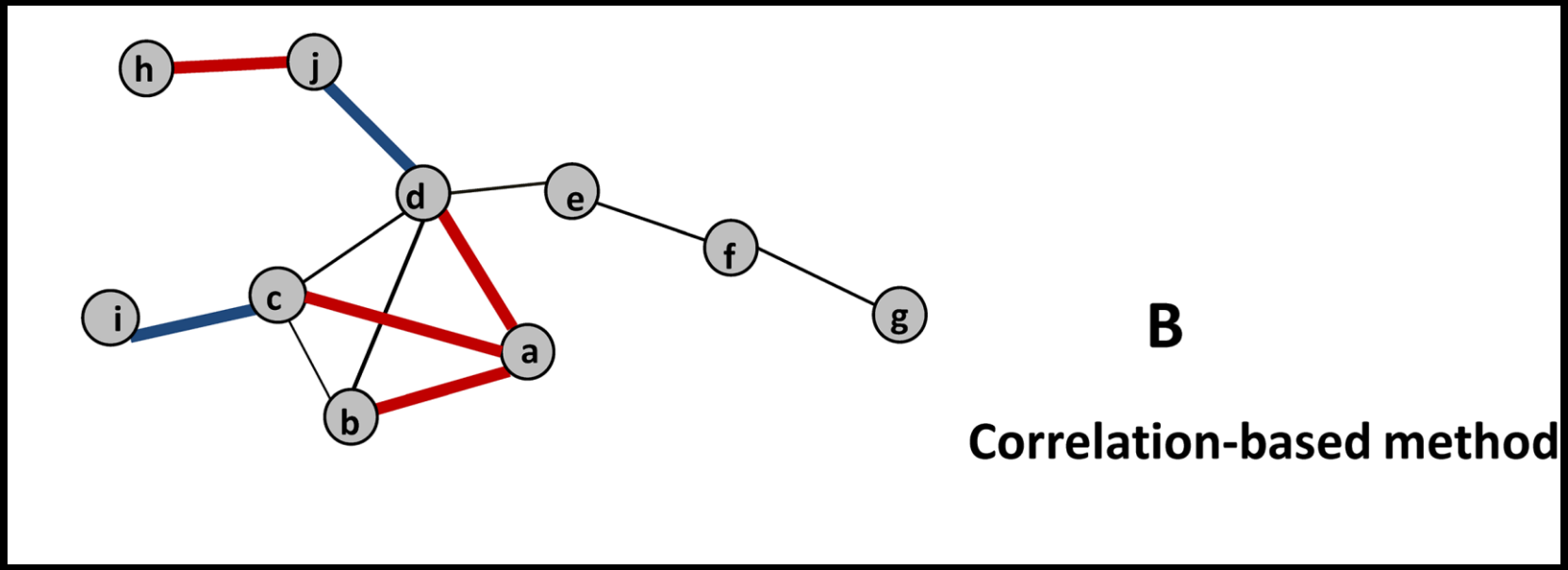
Results –

- Better discriminative power than single gene markers
- Increased reproducibility across two datasets



Subsequent improvement of the method:
Dao et al. Optimally discriminative subnetworks Bioinformatics 2011

GROUP 2: Correlation based methods



Idea

- Correlation in expression changes of two neighboring nodes is suggestive of joint function
- Identify modules with correlated changes

Example 1: Jointly active subnetworks Ulitsky et al 2007

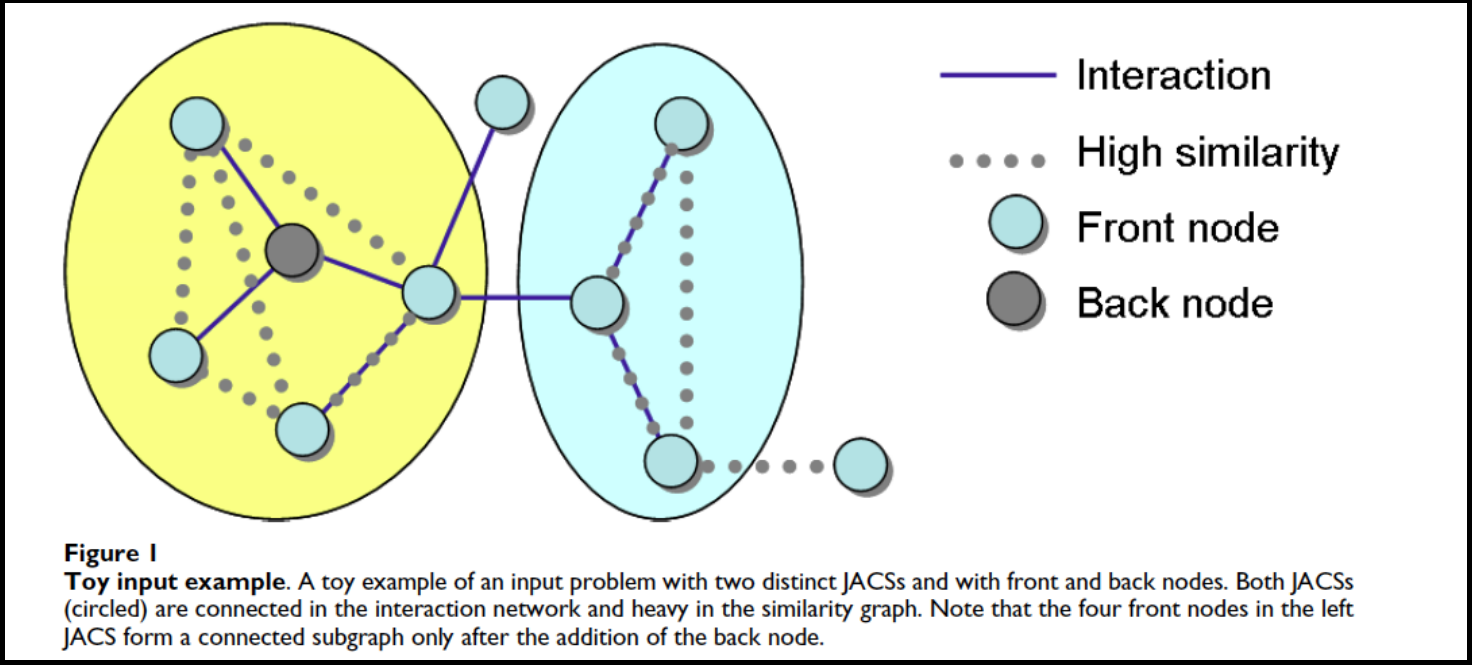


Figure 1
Toy input example. A toy example of an input problem with two distinct JACSs and with front and back nodes. Both JACSs (circled) are connected in the interaction network and heavy in the similarity graph. Note that the four front nodes in the left JACS form a connected subgraph only after the addition of the back node.

- detection of relatively small, high-scoring gene sets, or *seeds*:- for each node, the set consisting of it along with the neighboring nodes that are connected to it via positive-weighted edges
- seed improvement
- significance-based filtering - an empirical p-value cluster score calculated using sampling randomly gene groups of the same size.

Application: osmotic response in yeast

Example 2: IDEA

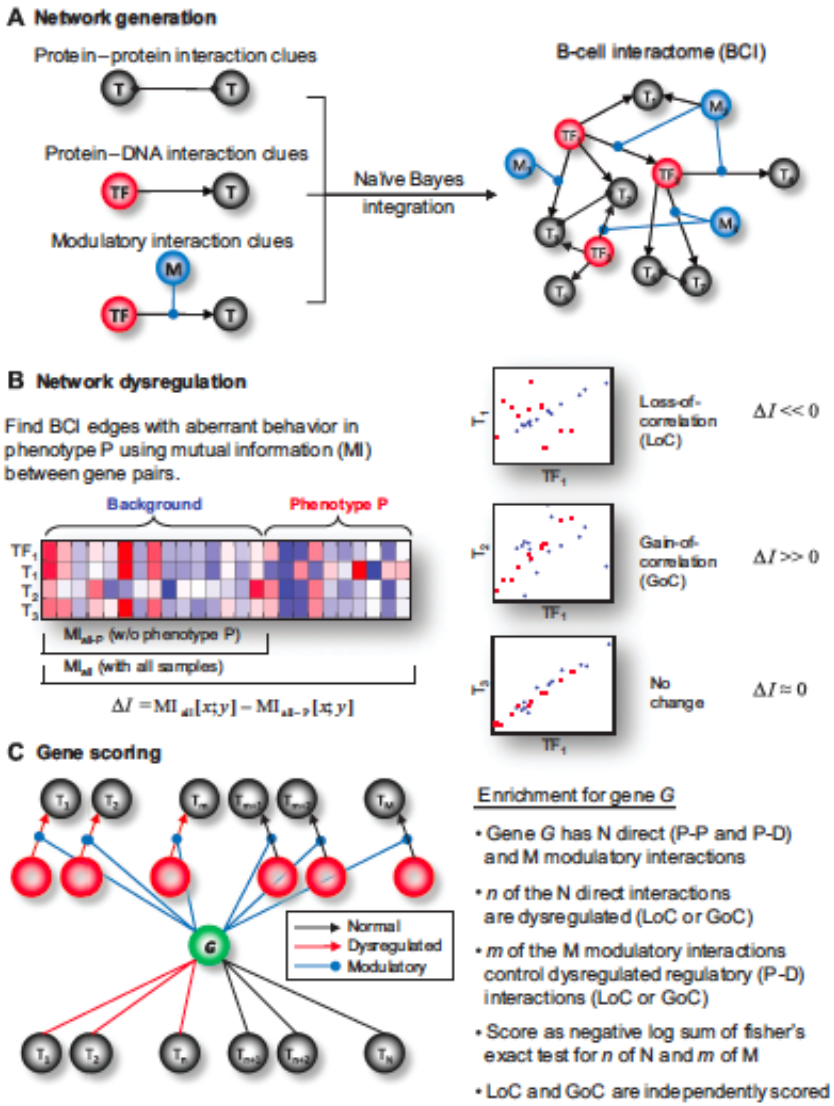
Interactome Dysregulation Enrichment Analysis

Mani et al. Molecular Systems Biology 2008

Identify perturbed network edges maximum loss/gain of correlation in disease state relative to a reference state

Underlying network – combined PPI, transcription, posttranscriptional modifications (here a predicted network is used)

Box 1 Interactome Dysregulation Enrichment Analysis (IDEA)



An overview of the proposed network-based analysis to characterize oncogenic mechanisms and pharmacological interventions. (**A**) In step 1, a comprehensive network of interactions is generated for B cells using a Bayesian evidence integration approach, including predictions of post-translational modifications. In this diagram, transcription factors are shown in red, non-transcription factors in gray, and modulators are shown in blue. Directed arrows indicate protein–DNA (P-D) interactions, and undirected indicate protein–protein (P-P) interactions or modulation events. Evidences, or clues, include curated databases, literate mining, orthologous interactions from model organisms, and reverse engineering algorithms. (**B**) In step 2, each interaction is analyzed to determine which show aberrant behavior in a specific phenotype (P); that is, interactions that show correlation in all samples except P (TF1 and T1), or interactions that are not correlated in any samples except P (TF1 and T2). These dysregulated interactions are classified as LoC or GoC, respectively, for every edge in the BCI. (**C**) In step 3, these dysregulated interactions are pooled together and a statistical enrichment is calculated which identifies genes having an unusually high number of these interactions in its neighborhood, either through direct or modulated links.

Application – B-cell lymphoma

Predicting oncogenes and perturbation targets in B-cell lymphoma
KM Mani *et al*

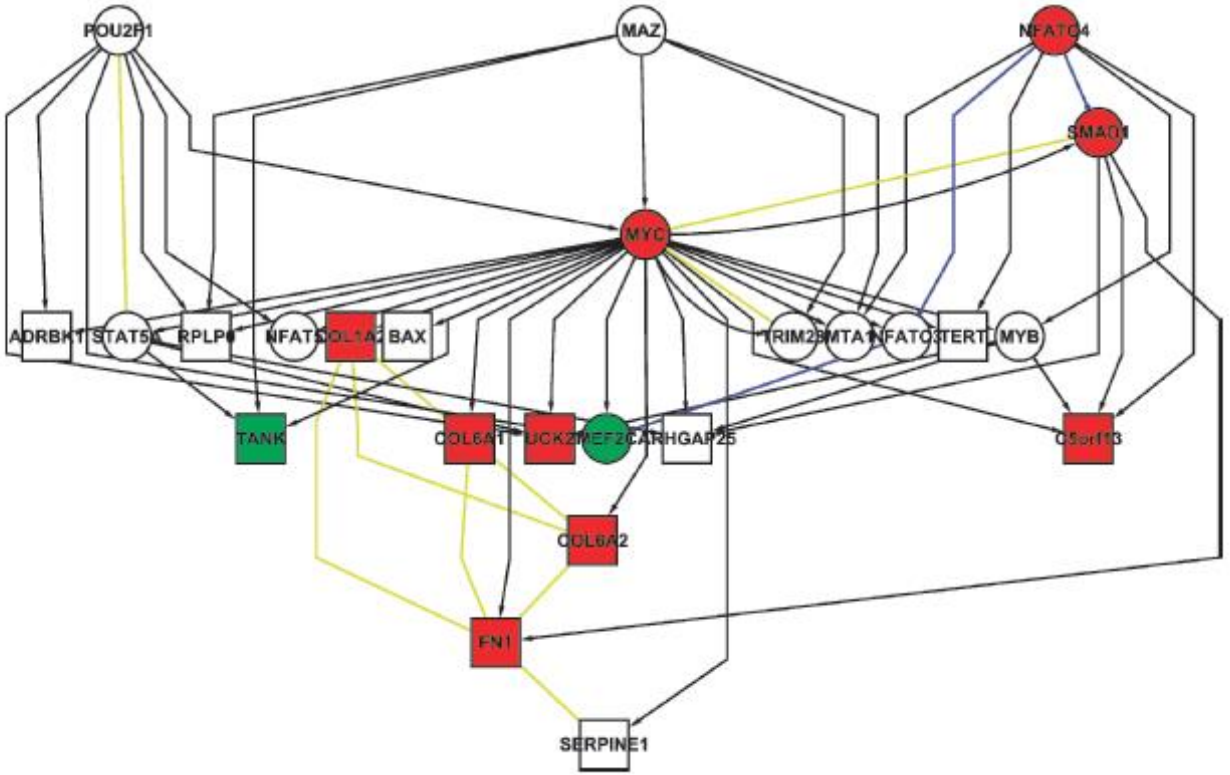
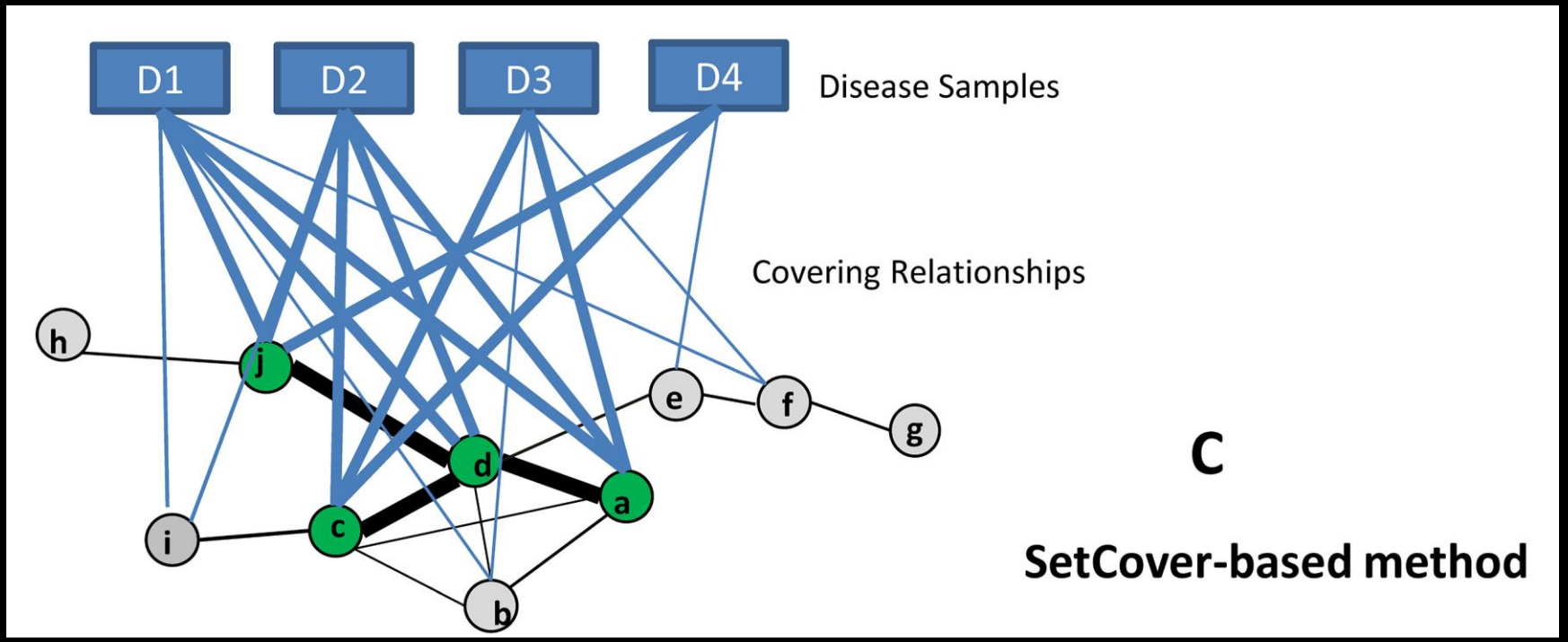


Figure 2 BL module: A network visualization of the top 25 scoring genes in BL. Transcription factors are shown as circles, whereas other proteins are shown as squares. Protein–protein interactions are also shown in beige, protein–DNA interactions are black with an arrowhead, and transcription factor-modulated interactions are shown in blue with a circular endpoint. Red/green indicates overexpression or underexpression ($P < 1e-8$), respectively in BL versus GC cells. There are some notable characteristics of this figure. First, all 25 genes form a connected module, which would not occur by chance. Second, *MYC* appears to be a central regulator of this module, as a full 21 out of the 25 members are *MYC* targets. *MYC* also appears regulated by *MYC*-associated zinc-finger protein (*MAZ*), which is also not differentially expressed. Third, there are interesting sets of genes that emerge, such as *SMAD1*, which is known to be associated with some NHL, and members of the *NFAT* family, including *NFATC3*, *NFATC4*, and *NFAT5* (these proteins are members of the Wnt-signaling pathway). There also appears to be a protein complex of *COL1A2*, *COL6A1* *COL6A2*, and *FN1*, which are all upregulated (and members of the cell signaling and ECM–receptor interaction pathway). These module diagrams can serve as a useful platform for further hypothesis generation and biochemical investigation.

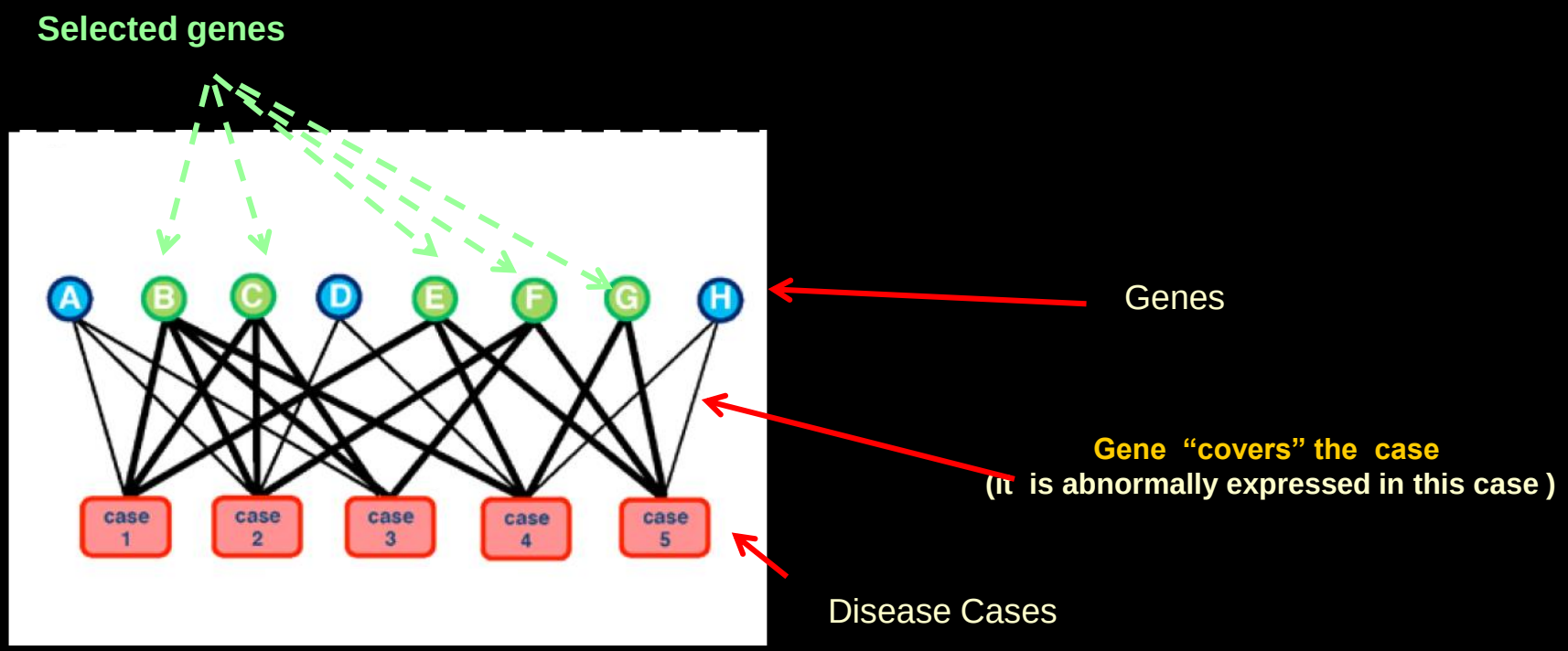
GROUP 3: Set cover based methods



Basic Idea

- Build a bipartite graph the two sets of bipartitions are diseases and genes
- Draw an edge between gene and disease case if gene is dysregulated in given case – “gene covers the case”
- Find gene subnetworks that are optimal with respect to selected optimization function and at the same time cover all disease cases
- In some applications each gene case is required to be covered multiple times

Covering with individual genes



Can be seen as a feature selection method:

- **Minimization of cover forces using genes dysregulated in many cases**
- **Each disease case covered min # times accounts for disease heterogeneity**

Kim et al. PloS CB 2011

DEGAS –de novo discovery dysregulated pathways in human diseases

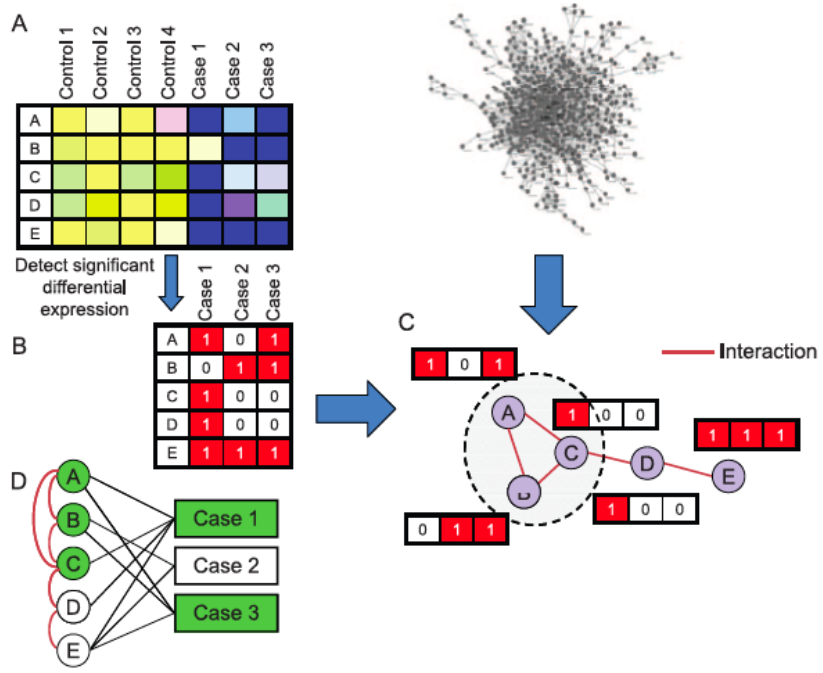


Figure 1. A dysregulated pathway (DP). (**A**) The input to our method consists of expression data of case and control cohorts and a protein interaction network. (**B**) The expression data are converted into a binary genes over cases matrix in which "1" appears in position (i,j) if gene i is dysregulated in case j (relative to the expression levels of i in the control cohort). (**C**) The interaction network. The vector next to each protein is the dysregulation status (0 or 1) of that gene in each case. A DP is a minimal subnetwork in which at least k genes are dysregulated in all but l cases. In the shown example, $k=2$ and $l=1$. In the circled subnetwork, two out of the three genes are dysregulated in the first and the third case (the second case is the outlier). (**D**) An alternative representation of the data in C, as a bipartite graph. Genes are on the left and cases are on the right. The blue edges are protein interactions and the gray edges connect the genes with cases in which they are dysregulated.
doi:10.1371/journal.pone.0013367.g001

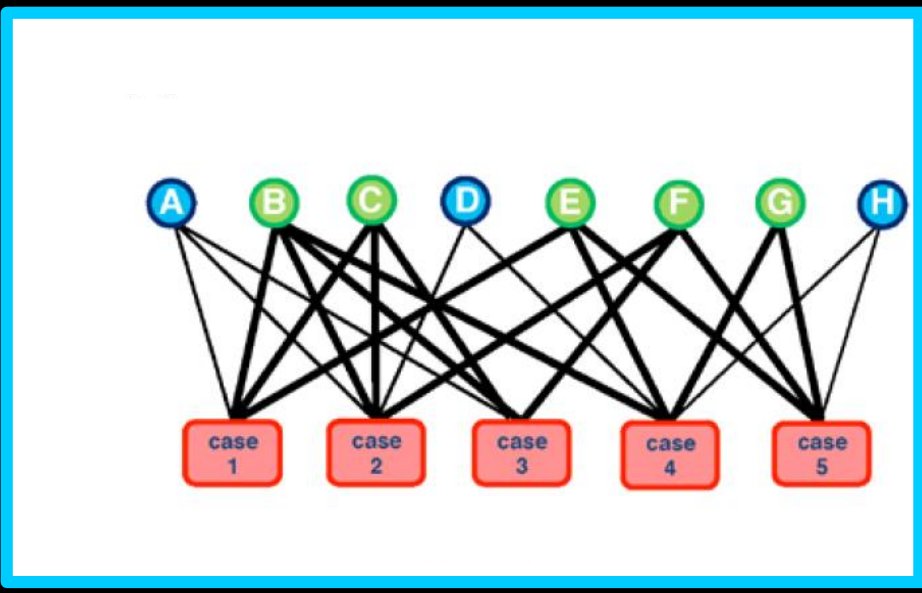
Idea:
Find ~~min-size~~ radius connected subnetwork covering all cases k-times

Diseases: Alzheimer's disease, Asthma, Helicobacter pylori infection
HD Huntington's disease, Parkinson among others

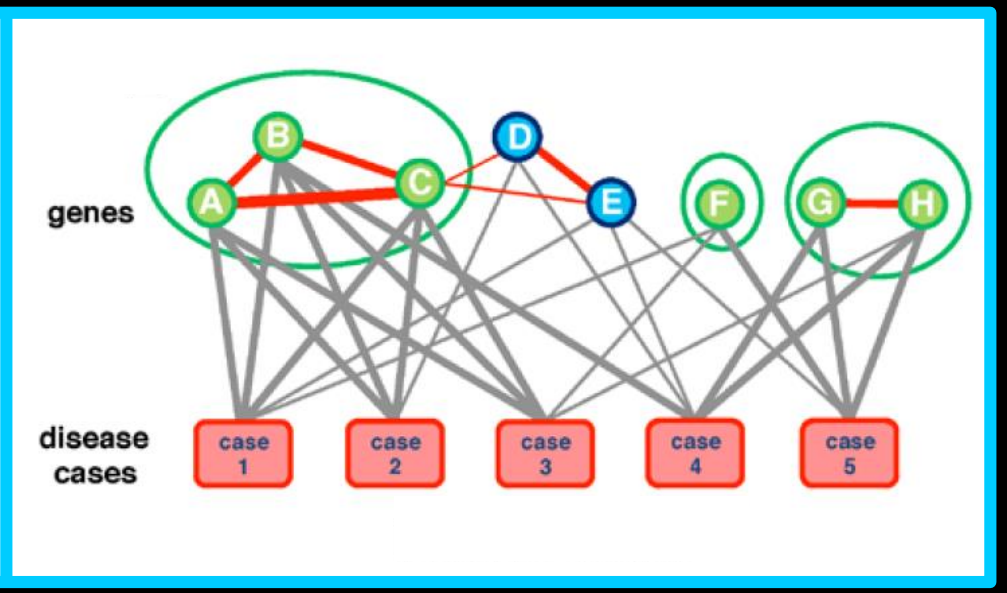
Ulitsky et al. PloS One 2010

Multi-module cover

Gene Cover



Module Cover



Optimization problem:

Find smallest number of genes so that each disease case is covered k-times

Kim et al. *PloS CB 2011*

Optimization problem:

Find smallest total cost modules so that each disease case is covered k-times

Kim et al. ***PSB 2013***

Cost function for this example

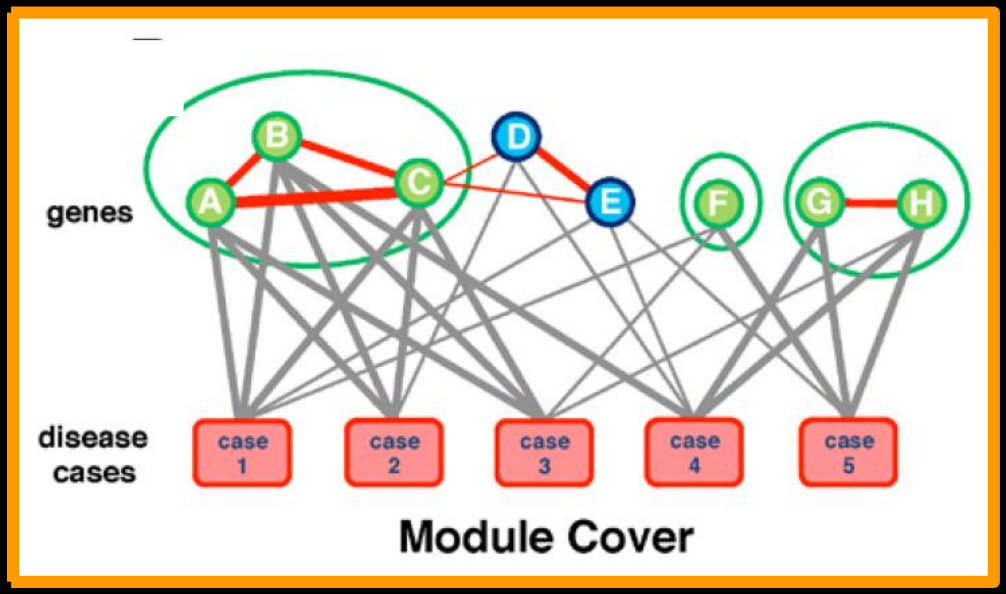
Recall optimization problem:

Find smallest cost set of modules so that each* disease case is covered at least k times

*Or each but a small number of outliers

Cost is a function of:

- ⬇ **Small distance in the network of genes in same module**
- ⬇ **Similar of some additional features**
- ⬆ **number of modules**



Kim et al. PSB 2013

Defining module cost

$$Cost(M) = \alpha + |M| - \sum_{x \in M} \sum_{y \in M, y \neq x} w(x,y)/(|M| - 1)$$

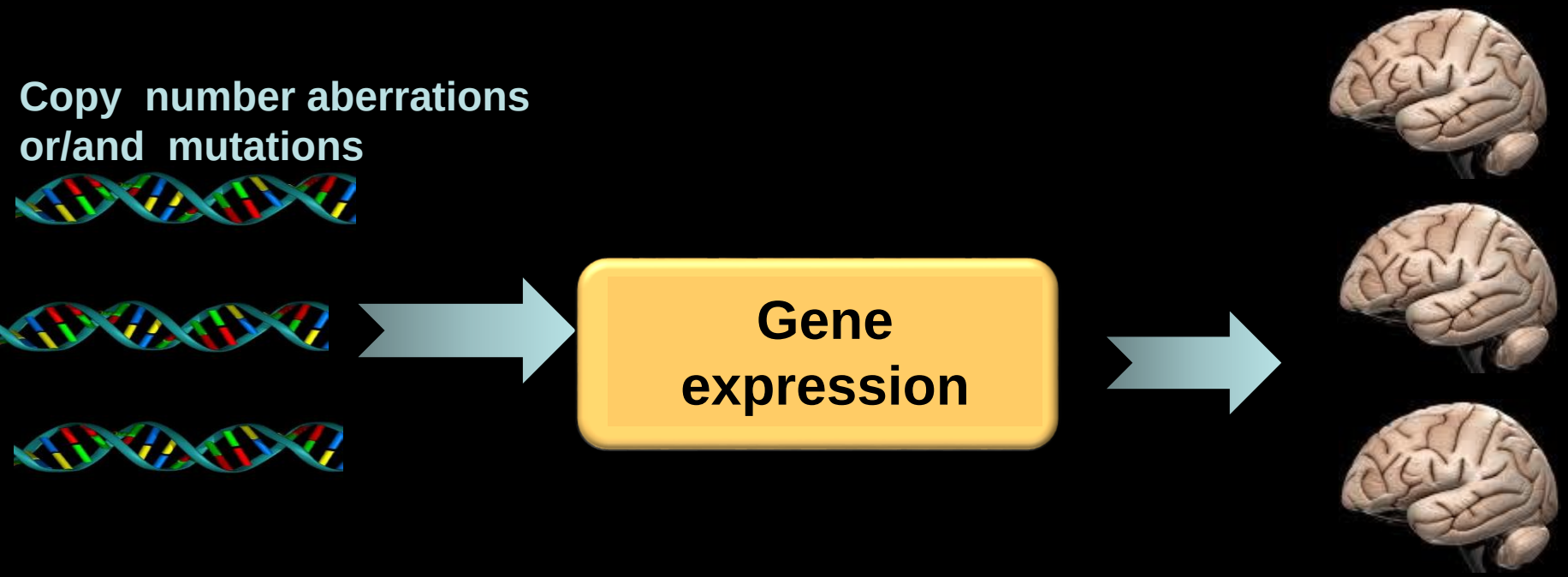
Cost for module initiation
Module size
Adjusted similarity – similarity + closeness in the network

$$adjusted\_sim(g_1, g_2) = sim(g_1, g_2)^{1+(distance(g1, g2)-1)/(avg\_dist\ -1)}$$

$$w(g_1, g_2) = adjusted\_sim(g_1, g_2) - \theta$$

Threshold parameter

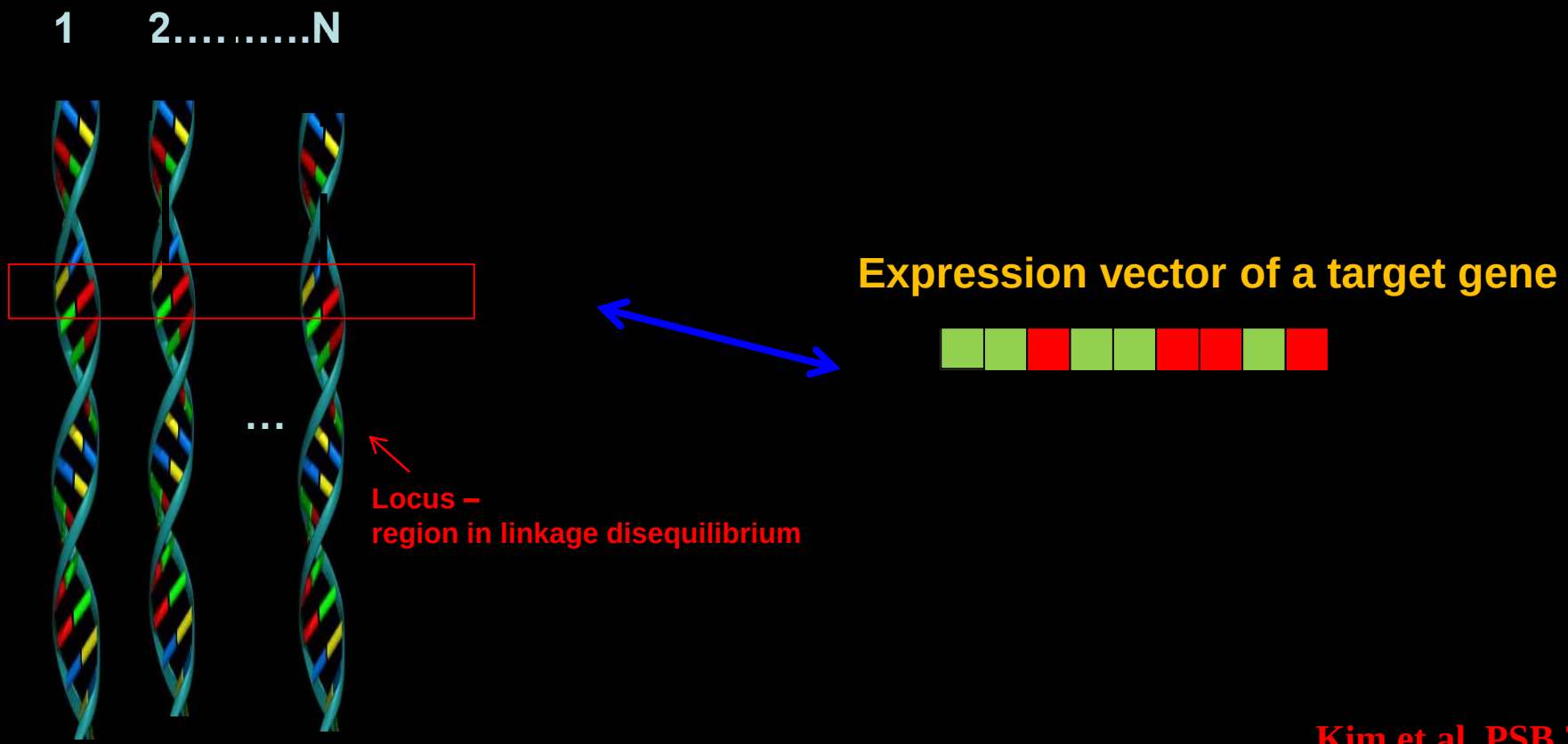
Sim socre depends on application: Example eQLT derived modules in giloblastoma multiforme



Find modules whose expression is jointly dysregulated by the same mutations/copy number variations

eQTL - expression Quantitative Trait Loci analysis

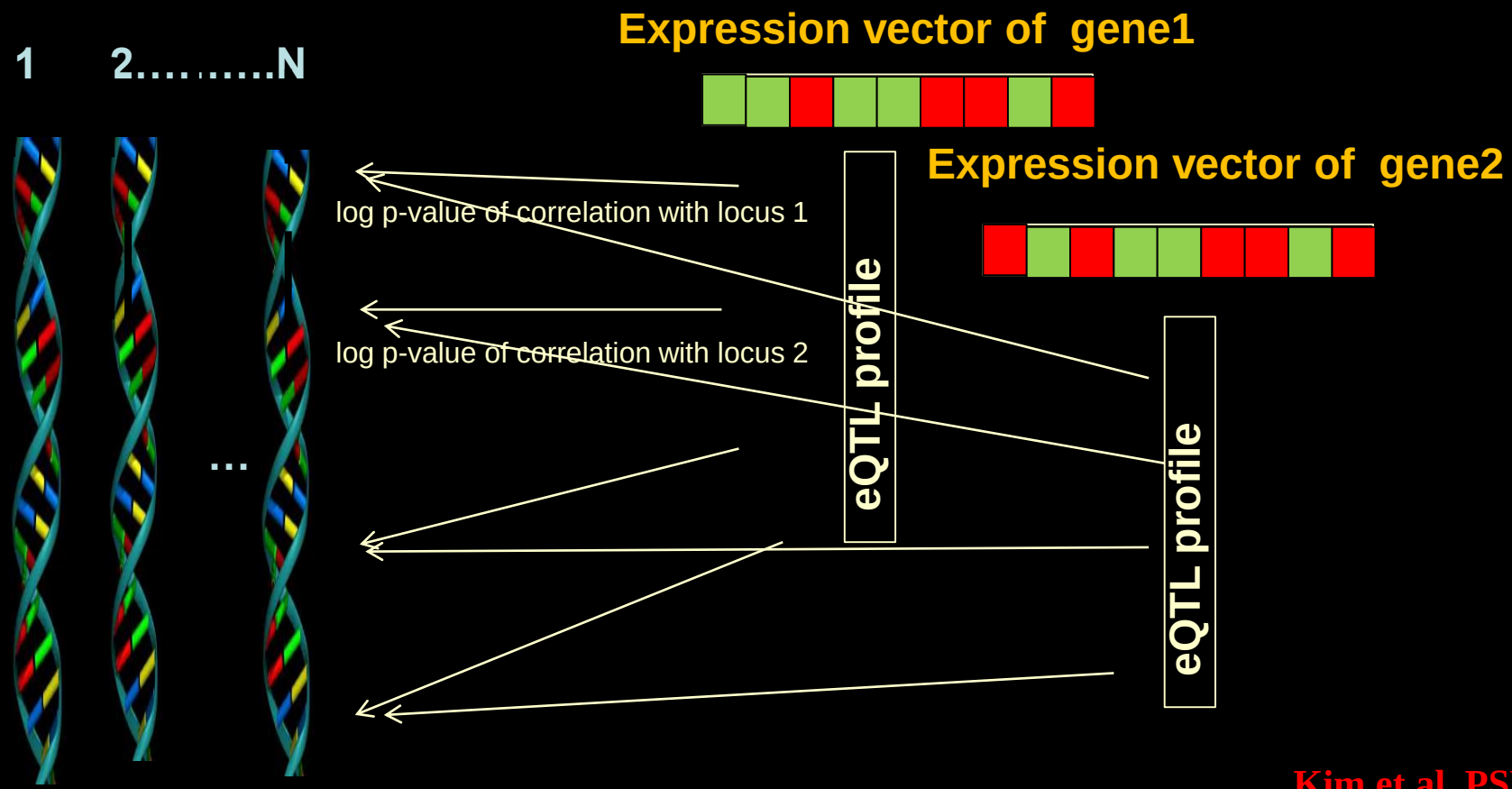
Correlation between expression of "target" gene and genetic variations of putatively causal loci



Kim et al. PSB 2013

Similarity of eQTL profiles

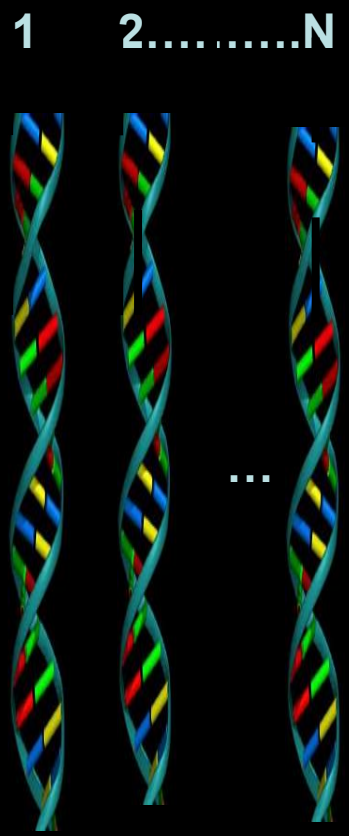
- **Goal – genes in a module should respond to the same genomic alterations**

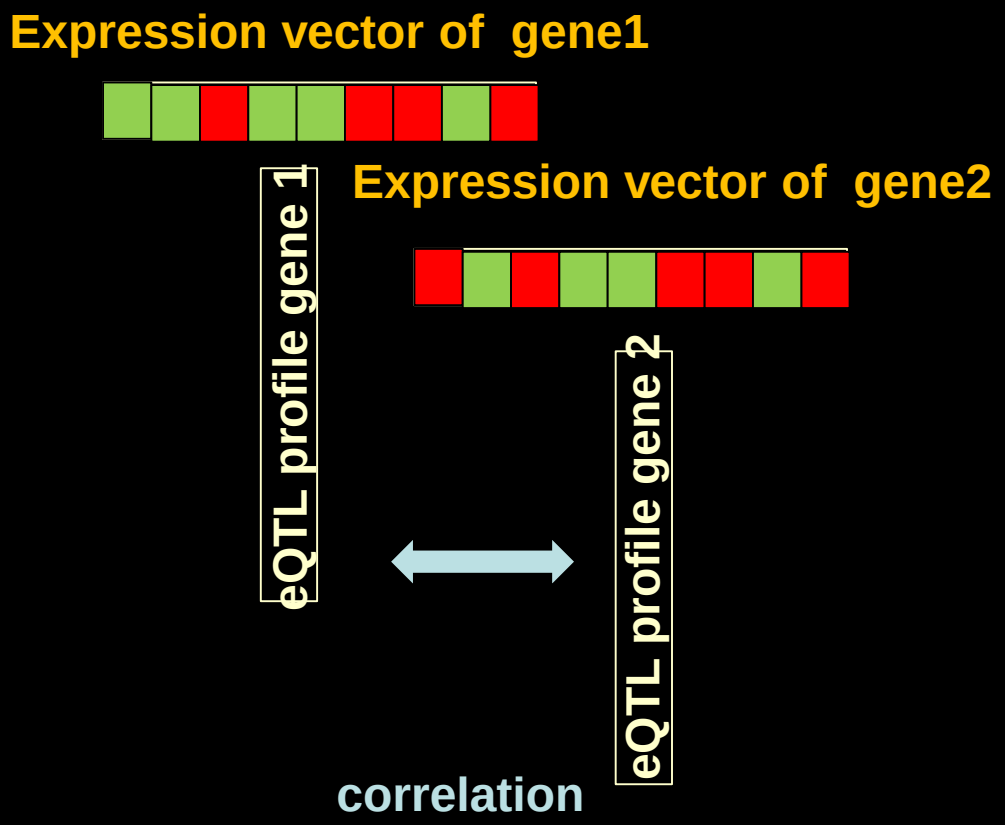


Kim et al. PSB 2013

Similarity of eQTL profiles

- Goal – genes in a module should respond to the same genomic alterations





Kim et al. PSB 2013

Cost function for this example

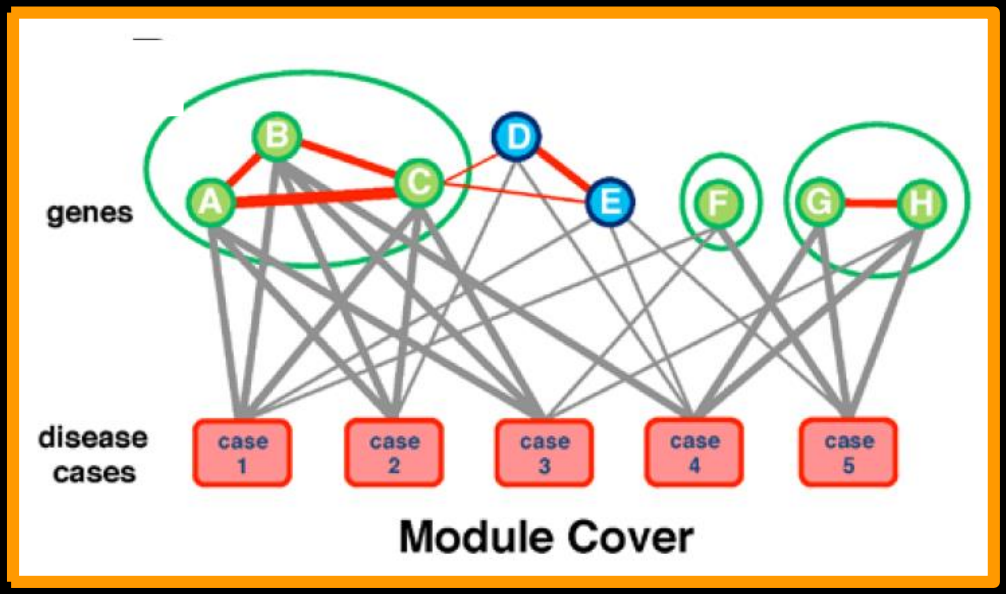
Recall optimization problem:

Find smallest cost set of modules so that each* disease case is covered at least k times

*Or each but a small number of outliers

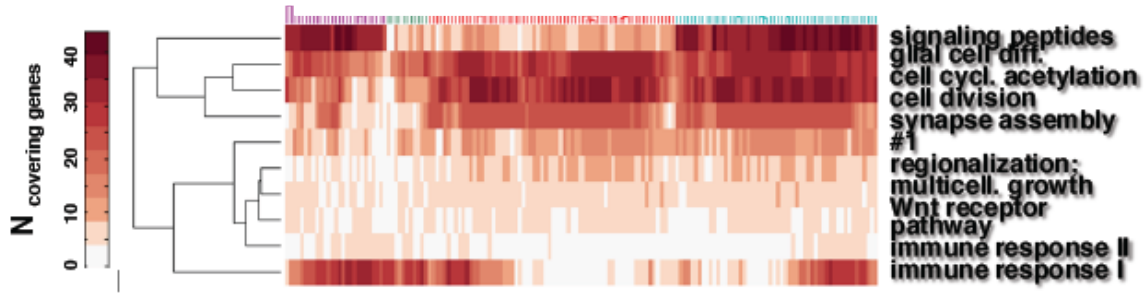
Cost is a function of:

- ⬇ **Small distance in the network of genes in same module**
- ⬇ **Similar eQTL profile**
- ⬆ **number of modules**

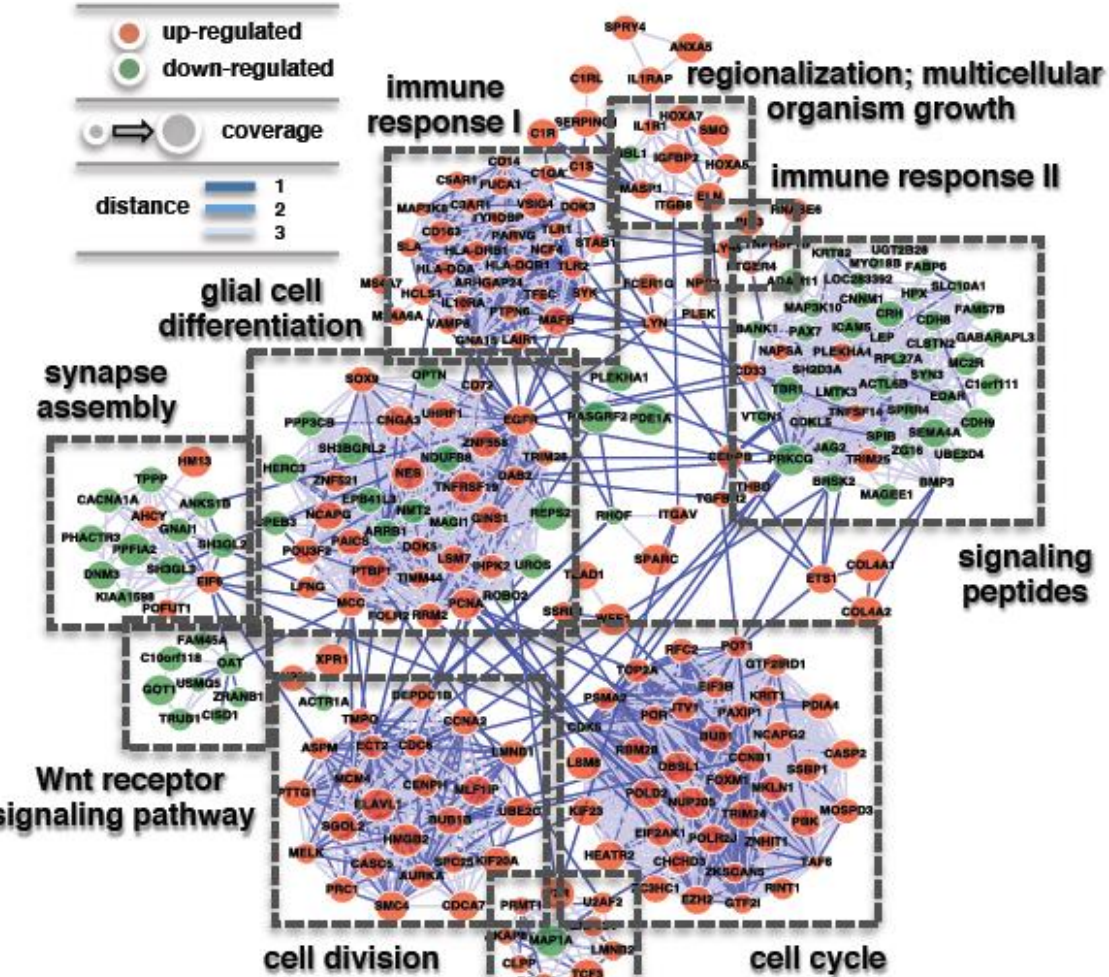


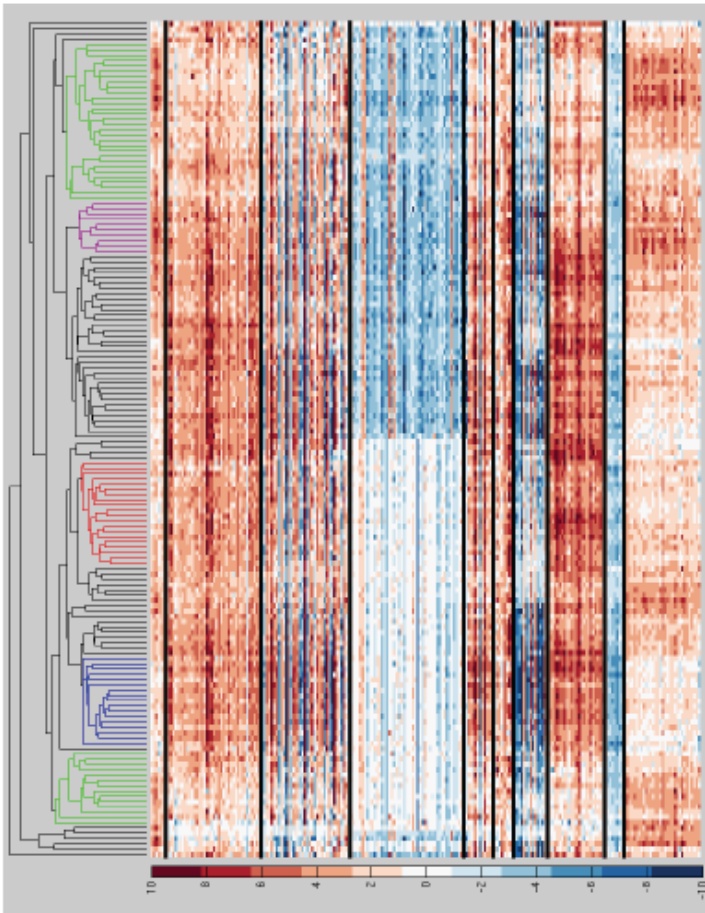
Kim et al. PSB 2013

GBM- Rembrandt



Immune Response II
Glial Cell Differentiation
Cell Cycle, acetylation
Signaling peptides
1 Regionalization; Multicell. Growth
Synapse Assembly
Cell Division
Wnt receptor signaling pathway
Immune Response I



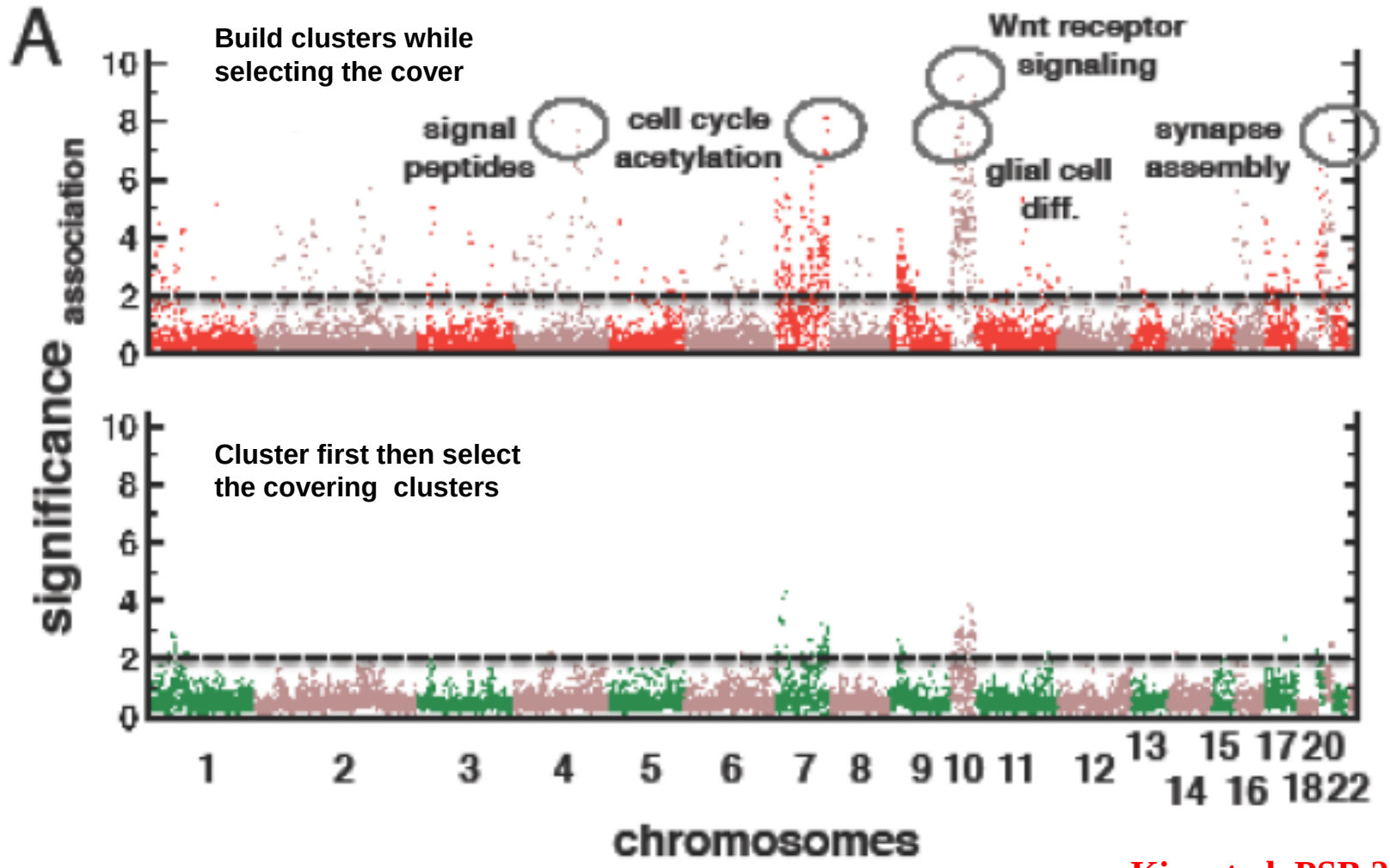


Is there an advantage to simultaneous module discovering and computing the cover?

Alternative approaches –

- **compute modules first (based on the similarity of interest) and then compute the cover using pre-defined modules**
- **Compute single node cover and then cluster**

Mapping back to causes shows advantage of simultaneous cover and clustering



Kim et al. PSB 2013

properties of jointly perturbed modules

- Contain genes abnormally expressed in at least some fraction of cancer cases
- Contain genes hypothesized to be jointly regulated by the same genetic aberrations
- Contain genes close to each other in the interaction space
- Allows for different samples might to be covered by different modules
- Each sample is covered by at least one module (in an alternative formulation all but a small number of samples are covered by some module)

Summary

- We looked at dysregulated modules using genotypic data, expression data and combination of the two data types. Method included
  - Scoring
  - Correlation
  - Set cover
- Advantages of module based approaches
  - Increased statistical power
  - Increased interpretability
  - Increased reproducibility
- Advantages of individual methods
  - Scoring – simplicity, can be used with smaller number of samples
  - Correlation – coherent modules
  - Set cover – dealing with heterogeneous data; can be combined with either two previous approaches by appropriate design of the optimization function